中文文本聚类常用停用词表对比研究*

官 琴 邓三鸿 王 昊
(南京大学信息管理学院 南京 210023)
(江苏省数据工程与知识服务重点实验室 南京 210023)

摘要:【目的】通过实验对比分析，比较不同停用词表对于不同类型的文本数据的作用效果，对停用词表的构建与使用提供参考意见。【方法】选取百度停用词表、哈尔滨工业大学停用词表以及四川大学机器智能实验室停用词表，基于三个不同语料库运用汉语分词技术、TF-IDF 特征评估函数以及 VSM 模型进行文本处理，并且采用 Java 编写的 K-means 算法进行聚类实验，通过准确率 P、召回率 R 和 F1 三个评价指标对不同聚类结果进行效果评估。【结果】不同停用词表对于不同类型的文本数据作用效果差异明显，词表的长度、内容结构是影响作用效果的直接因素，其中两字停用词作用效果最为明显。【局限】实验文本类型及数量有限，同时对于不同停用词表仅在词语数量及内容上做了简单的分析比较，未对停用词按照类别分类进行实验分析。【结论】停用词表对于文本聚类准确度有很大的影响，构建或选取适宜的中文停用词表极为重要。同时，过度增加停用词的数量并不会一直改善聚类结果。

关键词: 文本聚类 停用词 K-means
分类号: TP391

1 引 言

在信息迅猛发展的互联网时代，对庞大信息量的处理与利用使得文本挖掘这一技术受到广泛关注。1995 年，Feldman 等提出文本挖掘这个概念[1]，随后 Ahonen- Myka 等将数据挖掘技术直接用于经过预处理的文本信息，同时指出文本预处理对挖掘过程的效率至关重要[2]。文本预处理一般要占据文本挖掘大部分的时间，对于中文文本而言，这个过程包含中文分词处理，去停用词，特征提取以及空间向量表示这几个步骤，因此停用词的研究具有重要意义。

停用词最早起源于信息检索，Luhn 在信息检索的研究中发现部分词语出现频率很高但检索效果却较差[3]，他率先提出用噪声来表示这些词语[4]，即为停用词的雏形。在随后的研究中，有学者通过统计研究发现在英文文献中最常出现的 10 个词条的频次占一篇文本总词条频次的 20%-30%[5]，而 Frakes 等在信息检索的研究中认为在自动索引阶段提早考虑消除出现频率过高的词语可以提高检索速度，减少检索存储空间并且不会降低检索结果的准确性[6]，因此，Lo 等将停用词定义为经常出现在文本中但对信息检索没有帮助的应该消除的词语[7]，即，在基于词的检索系统中，停用词是指出现频率较高、没有太大检索意义的词，如“的、是、太、of”等[8]；在自动问答系统中，停用词因其问题的不同而动态变化[9]；在支持向量机的自动分类中则是指没有实际意义的虚词和类别色彩不强的中性词[10]；在文本挖掘中，停用词的判断更侧重于其是否能够表示文本特征。

停用词在文本处理过程中会存在很大的干扰性，不仅携带较少的文本信息，还会对其他词语产生一定的抑制作用，很大程度上影响文本处理效率和精准性。Yang 和 Pedersen 认为，将停用词按其出现的频数

通讯作者：邓三鸿，ORCID：0000-0002-6910-3935，E-mail：sanhong@nju.edu.cn。
*本文系中国地震局星火计划攻关项目“面向地震应急的空间智能决策方法研究”(项目编号：XH15019)和江苏省自然科学基金项目“面向专利预警的中文文本学习研究”(项目编号：BK20130587)的研究成果之一。

降序排列，用前 10 个停用词消减特征向量，不会产生负面影响；用前 100 个停用词消减特征向量产生的负面效果很小[11]。Silva 等也通过实验验证了去除停用词可以在很大程度上降低特征向量的维度并且提高文本分类的准确性[12]。因此，去除停用词在文本预处理过程中十分重要。

目前，可以通过构建停用词表去除停用词。停用词表有通用停用词表与专用停用词表之分，也有学者将停用词分为全停用词(True Full-stop Words)和半停用词(Semi-stop Words)[13]，其来源有人工构造与基于统计的自动学习两种方式[14]。Luhn 提出"词条的区分能力"这一概念，成为人工构造方法常用的判断标准[3]，Van Rijsbergen 利用统计学的方法构造出包含 250 个词条的停用词表[15]，Fox 也在 Brown Corous 的基础上统计分析出适用于普通英文文本的停用词表[16]。

基于统计的自动学习方法是指通过不断地标记筛选，从文本语料库中提取出高频词语，随后进行人工判定。现在较为成熟的停用词识别算法有：文本频率、词频统计、熵计算、CHI 统计等[17]。文献[18]提到一种依据联合熵选取停用词的方法；文献[19]提出一种基于统计和语言学结合的停用词选取方法； Lo 等设计了一种基于词条的随机抽样的抽取方法，并指出最有效的停用词表是经典停用词表与新方法自动抽取的停用表的融合[7]；Zou 等提出一种基于统计和信息论模型的停用词选取方法[20]；在中文情感分类中，也构建出 5 种包含不同词性的停用词表[21]。基于统计的自动学习方法已成为停用词表构建的主要方法，同时加以人工判定的辅助，并取得不错的效果。

当前，专业停用词表的研究也受到关注，如医学、化学、计算机等领域，主要通过对该领域大量文本进行检查分析，经过概率分析及内容分析予以提取[22]。但该方法具有一定的局限性，当文本分布不均时准确率不高。Makrehchi 等为此提出一种利用参数和输入比较敏感的分类器来判别停用词内容改变对分类结果的影响，从而确定停用词表的内容[23]。

英文停用词表的研究已取得一定成果，而中文停用词表由于起步时间较晚，目前深入的研究还较少，暂未得到广泛认可的停用词表。文献[24]分别比较去除不同词性词语的停用词表对中文情感分类的影响，发现使用去除形容词、副词及动词的停用词表效果最好，而传统的主题分类停用词表对于情感分类帮助不大，可以得出构建或选择精准的停用词表往往会起到事半功倍的效果。

目前主流的通用中文停用词表有百度停用词表、哈尔滨工业大学停用词表以及四川大学机器智能实验室停用词表，鉴于去除停用词对于分类所需的特征向量集及分类效果有很大的影响，因此本文主要目的在于利用多个不同的中文文本语料库，采用聚类算法对常用的停用词词表进行实验分析，旨在对不同停用词表的适用范围及使用效率进行比较研究，找出不同领域文本信息处理构建、选择和使用停用词词表的依据和准则。

2　实验过程准备

2.1　实验过程

本次实验由 4 部分组成，分别为文本收集、文本处理、聚类处理以及效果评估，具体流程如图 1 所示。

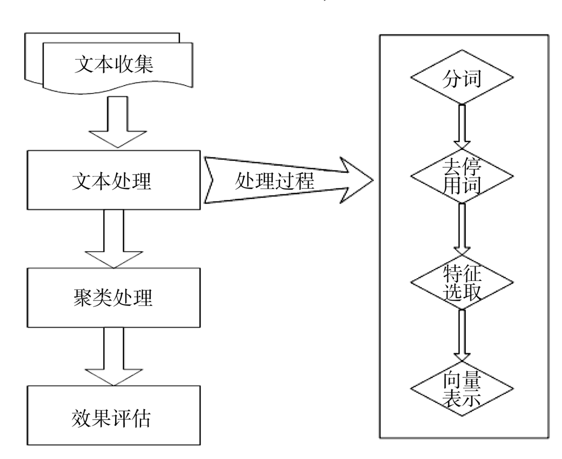


图 1　实验流程

实验采用三个不同的中文语料库，分别是搜狗实验室的搜狐新闻数据[25-27]，复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供的复旦文本语料库[22,28-29]，以及文献[30]提及的中文文本分类语料库[31]，这三个语料库的文本质量较高，类型齐全，覆盖面广。在文本处理过程中，分词处理时采用的是由中国科学院计算技术研究所开发的汉语词法分析系统 ICTCLAS；选取主流且应用范围较广的停用词表，分别为百度停用词表、哈尔滨工业大学停用词表以及四川大学停用词表[32-33]。同时在特征选取中采用 TF-IDF 评估函数，利用 VSM 模型对文本进行向量表

示。在聚类处理过程中，采用 Java 编写的较为简单高效的 K-means 算法，运用基于人工标准的评价方法，利用准确率 P(Precision)、召回率 R(Recall)以及 F1(F1-measure)三个指标对分类效果进行综合评价[16]。

实验文本数据从语料库中随机抽样选取，在各语料库中的经济、IT、军事、体育和艺术这 5 大类中抽取序号能被 5 整除的文本，并为其重新标号，每一个语料库各抽取 640 个文本，如表 1 所示。测试文本数为 1 920 个，每个文本至多属于一个类别；其中文本类型包括新闻、文献、文摘等，采用控制变量法进行实验测试，对实验结果准确统计记录，用于分析比较。

表 1　人工分类文本统计表

	经济	IT	军事	体育	艺术
文本数	110	164	76	150	140

2.2　停用词表内容分析

本文采用的三个停用词表基本情况如表 2 所示。

表 2　实验停用词表内容统计

停用词表	符号	英文	单字词	两字词	三字词	四字词	其他	共计
百度	7	547	173	620	29	19	0	1 395
四川大学	0	0	26	663	80	84	6	859
哈尔滨工业大学	236	0	167	290	23	19	0	750

可以看出，各停用词表的差异较大，百度停用词表包含部分单字符、英文停用词以及中文停用词，如“able”、“一”以及“不是”等，两字词比例较大；四川大学停用词表包含很多常见俗语及三字词、四字词，如“打开天窗说亮话”、“何乐而不为”以及“换言之”等，单字词数量相对较少；而哈尔滨工业大学停用词表则包含大量的中英文字符，如“*”、“Δ”以及“......”等。三个停用词表中，百度停用词表停用词数量高达 1 395 个，主要在于其包含了 547 个英文停用词；而两字词在三个停用词表中比例较高，其中四川大学停用词表包含 663 个两字词，显然是为了能保证最大程度匹配并去除停用词，因为在中文分词的结果中大部分都为两字词串[34]。

表 3 显示了三个停用词表的重合情况，可以看出，这三个停用词表的单字词、三字词及四字词的重合率很高，基本达到 80%以上，因此三个停用词表的区别主要体现在两字词上。其中百度停用词表和四川大学停用词表两字词数量相近，重合率约为 50%，在两字词上的差异较大；百度停用词表和哈尔滨工业大学停用词表有较高的词语重合度，其主要差异在于哈尔滨工业大学停用词表包含的两字词较少；而这三个表共有的停用词数量有 337 个。在实验中，笔者合并了三个停用词表作为一个新的停用词表(命名为全停用词)以测试停用词表的长度是否也会影响文本聚类效果。

表 3　停用词表重合词条统计

对比词表	单字词	两字词	三字词	四字词	共计
百度–四川大学	22	311	23	19	374
百度–哈尔滨工业大学	167	288	22	18	493
四川大学–哈尔滨工业大学	22	276	22	18	338
百度–四川大学–哈尔滨工业大学	22	275	22	18	337

3　实验结果展示

本次实验分别固定待比较的停用词词表，将搜狗语料库、复旦文本语料库和中文文本语料库作为实验数据进行文本聚类处理，统计每一簇中各类型文本数量，以其数量最多的一类作为该簇的正确文本类型，同时进行评价指标 P、R、F1 的计算。其中表 4 至表 6 为三个停用词表处理后的文本聚类结构，表 7 是将百度停用词表之中的英文停用词去掉后得到的结果，而表 8 则是应用全停用词表得到的聚类结果。

表 4　百度停用词表实验结果统计

	指标	第一簇(艺术)	第二簇(经济)	第三簇(体育)	第四簇(IT)	第五簇(军事)	平均值
复旦语料库	P	0.924	0.965	0.930	0.763	0.608	0.838
	R	0.964	1	0.440	0.963	0.816	0.837
	F1	0.944	0.982	0.597	0.851	0.697	0.814
	指标	第一簇(艺术)	第二簇(体育)	第三簇(经济)	第四簇(军事)	第五簇(IT)	平均值
搜狗语料库	P	0.739	0.693	0.615	0.477	0.521	0.609
	R	0.929	0.813	0.582	0.553	0.445	0.664
	F1	0.823	0.748	0.598	0.512	0.480	0.632
	指标	第一簇(艺术)	第二簇(体育)	第三簇(经济)	第四簇(IT)	第五簇(军事)	平均值
中文语料库	P	0.882	0.803	0.831	0.817	0.233	0.547
	R	0.964	0.600	0.936	0.652	0.368	0.704
	F1	0.921	0.687	0.884	0.725	0.285	0.700

表 5　四川大学停用词表实验结果统计

	指标	第一簇(艺术)	第二簇(经济)	第三簇(体育)	第四簇(IT)	第五簇(军事)	平均值
复旦语料库	P	0.907	0.957	0.971	0.963	0.432	0.846
	R	0.979	1	0.447	0.976	0.789	0.838
	F1	0.942	0.978	0.612	0.969	0.558	0.812
	指标	第一簇(体育)	第二簇(艺术)	第三簇(经济)	第四簇(IT)	第五簇(军事)	平均值
搜狗语料库	P	0.614	0.424	0.743	0.455	0.040	0.455
	R	0.847	0.443	0.555	0.305	0.053	0.441
	F1	0.712	0.434	0.640	0.365	0.046	0.439
	指标	第一簇(艺术)	第二簇(体育)	第三簇(经济)	第四簇(IT)	第五簇(军事)	平均值
中文语料库	P	0.899	0.993	0.644	0.839	0.326	0.740
	R	0.950	0.893	0.791	0.634	0.421	0.738
	F1	0.924	0.940	0.701	0.740	0.367	0.734

表 6　哈尔滨工业大学停用词表实验结果统计

	指标	第一簇(艺术)	第二簇(经济)	第三簇(体育)	第四簇(IT)	第五簇(军事)	平均值
复旦语料库	P	0.924	0.948	0.943	0.732	0.667	0.843
	R	0.950	1	0.440	0.970	0.816	0.835
	F1	0.937	0.973	0.600	0.834	0.734	0.816
	指标	第一簇(艺术)	第二簇(体育)	第三簇(经济)	第四簇(IT)	第五簇(军事)	平均值
搜狗语料库	P	0.788	1	0.545	0.535	0.092	0.592
	R	0.929	0.840	0.382	0.604	0.105	0.572
	F1	0.853	0.913	0.449	0.567	0.098	0.576
	指标	第一簇(艺术)	第二簇(经济)	第三簇(军事)	第四簇(IT)	第五簇(体育)	平均值
中文语料库	P	0.937	0.438	0.135	0.743	0.605	0.572
	R	0.950	0.636	0.105	0.793	0.393	0.575
	F1	0.943	0.504	0.116	0.767	0.426	0.551

表 7　百度去英文停用词表实验结果统计

	指标	第一簇(艺术)	第二簇(经济)	第三簇(体育)	第四簇(IT)	第五簇(军事)	平均值
复旦语料库	P	0.964	1	0.440	0.963	0.803	0.834
	R	0.925	0.965	0.923	0.763	0.598	0.835
	F1	0.944	0.982	0.6	0.851	0.686	0.813
	指标	第一簇(艺术)	第二簇(体育)	第三簇(经济)	第四簇(军事)	第五簇(IT)	平均值
搜狗语料库	P	0.929	0.813	0.582	0.539	0.451	0.663
	R	0.739	0.924	0.615	0.465	0.528	0.654
	F1	0.823	0.865	0.598	0.499	0.486	0.654
	指标	第一簇(艺术)	第二簇(体育)	第三簇(经济)	第四簇(IT)	第五簇(军事)	平均值
中文语料库	P	0.964	0.600	0.936	0.659	0.382	0.708
	R	0.882	0.804	0.831	0.824	0.241	0.716
	F1	0.921	0.687	0.88	0.732	0.300	0.704

表 8　全停用词表实验结果统计

	指标	第一簇(艺术)	第二簇(经济)	第三簇(体育)	第四簇(IT)	第五簇(军事)	平均值
复旦语料库	P	0.938	0.965	0.943	0.976	0.438	0.852
	R	0.979	1	0.440	0.976	0.842	0.847
	F1	0.986	0.982	0.600	0.976	0.576	0.855
	指标	第一簇(艺术)	第二簇(经济)	第三簇(军事)	第四簇(体育)	第五簇(IT)	平均值
搜狗语料库	P	0.787	0.325	0.468	0.876	0.446	0.580
	R	0.871	0.245	0.789	0.567	0.482	0.591
	F1	0.827	0.279	0.588	0.688	0.463	0.586
	指标	第一簇(艺术)	第二簇(体育)	第三簇(经济)	第四簇(IT)	第五簇(军事)	平均值
中文语料库	P	0.882	0.833	0.831	0.831	0.467	0.769
	R	0.964	0.600	0.936	0.646	0.750	0.779
	F1	0.922	0.698	0.880	0.727	0.576	0.774

4　实验结果分析

以上实验结果数据均为直接统计计算所得，为了更好地对三个停用词表进行分析对比，将实验数据进行整合，分别从不同停用词表对同一文本类型(经济、军事等)的作用效果，以及不同停用词表对同一语料库的作用效果进行比对。

4.1　文本领域分析

本次实验数据涉及经济、IT、军事、体育和艺术这 5 个领域。在固定停用词表的情况下，以三个语料库作为实验数据，得到不同的聚类结果，分别挑选出不同语料库中各类型文本的 F1 值并求得三者的平均值，结果如表 9 和表 10 所示。

通过表 10 可以看出：

(1) 就这 5 个领域的文本类型来说，艺术类的聚类效果最佳，而军事类的聚类效果较差，二者的 F1 指标值相差近 50%，这与军事类文本数量较少有关，因此在进行聚类实验时，文本要保证一定的数量，才能够提取尽可能准确的特征值，构建更为精准的特征向量，避免在后续实验分析中带来干扰；

(2) 百度停用词表整体作用效果较好，F1 指标值高出其他两个停用词表 0.049 和 0.069，而其在经济以及军事领域表现相对突出。百度停用词表中最为突出的是其拥有较多的两字停用词，尽管其在数量上与四川大学停用词表不相上下，但其作用效果远高于四川

表 9　各语料库与文本领域综合统计表-F1 值

文本类型	语料库	百度	四川大学	哈尔滨工业大学
经济	复旦语料库	0.982	0.978	0.973
	搜狗语料库	0.598	0.640	0.449
	中文语料库	0.884	0.701	0.504
	平均值	0.821	0.773	0.642
IT	复旦语料库	0.851	0.969	0.834
	搜狗语料库	0.480	0.365	0.567
	中文语料库	0.725	0.740	0.767
	平均值	0.685	0.691	0.722
军事	复旦语料库	0.697	0.558	0.734
	搜狗语料库	0.512	0.046	0.098
	中文语料库	0.285	0.367	0.116
	平均值	0.498	0.324	0.316
体育	复旦语料库	0.597	0.612	0.600
	搜狗语料库	0.748	0.712	0.913
	中文语料库	0.687	0.940	0.426
	平均值	0.677	0.755	0.646
艺术	复旦语料库	0.944	0.942	0.937
	搜狗语料库	0.823	0.434	0.853
	中文语料库	0.921	0.924	0.943
	平均值	0.896	0.767	0.911

表 10　各领域文本聚类平均 F1 值

领域 / F1 / 停用词表	经济	IT	军事	体育	艺术	平均值
百度	**0.821**	0.685	**0.498**	0.677	0.896	**0.716**
四川大学	0.773	0.691	0.324	**0.775**	0.767	0.667
哈尔滨工业大学	0.642	**0.722**	0.316	0.646	**0.911**	0.647
平均值	0.745	0.699	**0.379**	0.699	**0.858**	0.676

大学停用词表，因此百度停用词表拥有较高质量的两字停用词，而两字停用词对文本聚类的作用效果的影响最为重要，所以在构建新的停用词表时要尽可能多地考虑两字停用词；

(3) 哈尔滨工业大学停用词表在 IT 类以及艺术类文本聚类中表现突出，而四川大学停用词表在体育类文本聚类中作用效果最好；观察二者内容结构发现，哈尔滨工业大学停用词表中包含其他停用词表较少含有的中英文字符，而四川大学停用词表中包含较多的三字、四字停用词，这些不同的特征是其作用效果不同的主要原因；

(4) 在此对不同停用词表适用的领域进行了实验分析，因此，在构建各领域专用停用词表时，可以依据其不同的表现进行选择参考。

4.2　不同语料库分析

将同一语料库中不同文本类型的 F1 值求和取平均值，得到固定停用词表对各语料库所产生的不同聚类效果，整合结果如表 11 和图 2 所示。

表 11　各语料库文本聚类效果平均值

语料 / F1 / 停用词表	复旦	搜狗	中文	平均值
百度	0.814	0.632	0.700	0.715
四川大学	0.812	0.439	0.734	0.662
哈尔滨工业大学	0.816	0.576	0.551	0.648

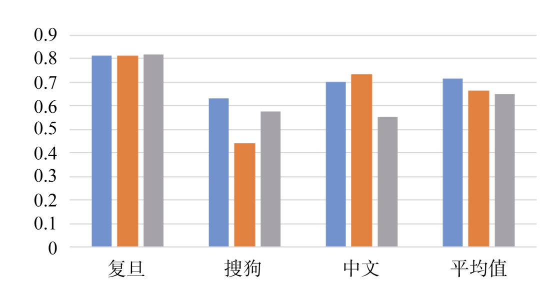


图 2　停用词表对于不同语料库作用效果对比

可以看到，百度停用词表对于搜狗语料库的作用效果较好，四川大学停用词词表比较适用于中文文本语料库，而哈尔滨工业大学停用词表更适合于复旦文本语料库。然而同一停用词表会对不同语料库产生较大差异的聚类效果主要还是取决于语料库中的文本类型，如经济、体育等，以及该语料库中文本的主要形式，文本类型在上文中已经讨论过，将不再考虑。

观察这三个语料库，可以看出，复旦语料库主要由大量的文献期刊组成，包含少量新闻报道评论；搜狗语料库均为各门户网站的新闻报道；而中文文本语料库中既有文献，又有新闻报道，还有一些邮件，组成较为复杂。初步得出结论这三个停用词表对文献期刊类文本的作用效果较好，而哈尔滨工业大学更胜一筹，对于新闻报道类的文本，百度停用词表的优势较大，四川大学停用词表作用效果较差，其更适合邮件

文献等类型的文本。

结合 4.1 节与 4.2 节实验，笔者进行如下总结：

(1) 去停用词作为文本处理的中间环节，具有至关重要的作用，其上一环节为中文分词，分词的结果关乎停用词的匹配，如“近年来”，可以切分为“近年”和“来”，也可直接切分为“近年来”，若按第一种切分方法，这三个停用词表均可顺利去除停用词，而若按照第二种则仅有四川大学停用词表可顺利去除，因为其含有这个三字停用词，因此，在构建停用词表时，要考虑到尽可能多的情况，才能够最大程度地匹配不同分词方法的处理结果；

(2) 去除停用词后的步骤为特征向量提取，而去除停用词也是为了在最大程度表现文本主题的前提下去除无用词语，降低特征向量的维度。不同停用词表对于不同语料库的作用效果不同，主要在于各停用词表构建时采用的是不同的语料库，因此，在针对不同语料库进行文本聚类时，选取与该语料库来源相近的停用词表会取得更好的效果，由此推广，在构建专用停用词时，要选取包含大量领域文本的语料库。

4.3　百度停用词表去除英文停用词结果分析

由于百度停用词表中包含大量的英文单词，而实验语料库均为中文文本，因此本次实验过程中，去除停用词表中的英文单词，并与原实验结果进行对比，结果如表 12 所示。

表 12　百度停用词表对比结果

文本类型	语料库	百度	百度(去英文)
经济	*复旦语料库*	*0.982*	*0.982*
	搜狗语料库	*0.598*	*0.598*
	中文语料库	0.884	0.880
IT	*复旦语料库*	*0.851*	*0.851*
	搜狗语料库	**0.480**	**0.486**
	中文语料库	**0.725**	**0.732**
军事	复旦语料库	0.697	0.686
	搜狗语料库	0.512	0.499
	中文语料库	**0.285**	**0.300**
体育	**复旦语料库**	**0.597**	**0.600**
	搜狗语料库	**0.748**	**0.865**
	中文语料库	*0.687*	*0.687*
艺术	*复旦语料库*	*0.944*	*0.944*
	搜狗语料库	*0.823*	*0.823*
	中文语料库	*0.921*	*0.921*

从表 12 可以看出，斜体标注的几个语料库聚类的结果不受是否去除英文停用词的影响，原因是这几个语料样本基本不含英文词汇。而加粗显示的几个语料库聚类的结果在去除英文停用词后略有上升，其他几个语料样本的聚类结果则略有下降，分析其文本，原因在于以 IT 类文本为例，经常会出现计算机领域的单词或字母，可以作为特征向量用于表征文本，如果去除则会降低聚类效果的准确性；而通过观察军事类文本，其中部分文本包含无用的英文单词，去除之后使得特征向量更加准确，提升了聚类效果。总体看来，这些样本中英文词汇的比例较低，且经常在文本中有特指意义，建议在去除停用词时无需考虑英文停用词的去除。

4.4　全停用词表结果分析

将三个不同停用词表整合为一个全停用词表，并将其用于聚类实验，同时，抽取三个停用词表聚类效果的最优值与之进行对比，结果如表 13 所示。

表 13　全停用词表对比结果

文本类型	语料库	最优值	全停用词表
经济	复旦语料库	0.982	0.982
	搜狗语料库	0.598	0.279
	中文语料库	0.884	0.880
	平均值	0.821	0.713
IT	复旦语料库	0.834	0.976
	搜狗语料库	0.567	0.463
	中文语料库	0.767	0.727
	平均值	0.722	0.722
军事	复旦语料库	0.697	0.576
	搜狗语料库	0.512	0.588
	中文语料库	0.285	0.576
	平均值	0.498	0.580
体育	复旦语料库	0.612	0.600
	搜狗语料库	0.712	0.688
	中文语料库	0.940	0.698
	平均值	0.755	0.622
艺术	复旦语料库	0.937	0.986
	搜狗语料库	0.853	0.827
	中文语料库	0.943	0.922
	平均值	0.911	0.912

可以看出，融合了三个停用词表的全停用词表聚类效果相比单个停用词表的聚类效果提升较大，但将其与各停用词表聚类效果最佳的文本类型结果相比

较，其优势并不明显，仅在军事类文本中提升了0.082，而在经济及体育类文本中下降了 0.108 和0.133，下降幅度较为明显，其余类型基本保持不变。事实说明，停用词表并不是包含的停用词越多越好，而是具有针对性比较好，并且能够针对文本已有信息，如文本来源、文本可能包含的类型去优化停用词表。

5 结　语

通过对三个停用词表具体内容进行比较分析，可以看出各停用词表的差异较为明显，主要体现在两字词语上，其在内容及数量上有显著差异，这三者的不同源于其源语料库及应用范围不同；而在具体的实验分析中发现，不同停用词表的使用对于聚类效果影响的差异是较为显著的，综合比较，百度停用词表对于三个语料库的平均作用效果最佳，去除英文词的百度停用词表聚类效果略有提升。三个不同停用词表对于艺术类文本的作用效果均高于其他类别，而对于军事类均没有起到很好的效果，且全停用词表并不能在聚类过程中取得最佳的效果，反而会在一定程度上降低聚类的精准度。因此，在处理具体聚类任务的时候，选取准确适合的停用词表是十分重要的，如果能够按需构建专业停用词表或者构建出更为全面的通用停用词表，效果会更佳，这也是日后研究的主要方向。在本次对比实验中，也存在一些不足，如实验文本数量较少，可能会导致实验结果的偶然性；聚类方法单一，仅采用 K-means 算法，不能排除算法对实验结果的影响；而对于停用词表作用的研究，未考虑按照停用词的类别进行分类处理，在后续的研究中将针对这些不足加以改进。

参考文献：

[1] Feldman R, Dagan I. Knowledge Discovery in Textual Databases (KDT)[C]//Proceedings of International Conference on Knowledge Discovery and Data Mining. 1995: 112-117.

[2] Ahonen-Myka H, Heinonen O, Klemettinen M, et al. Applying Data Mining Techniques in Text Analysis[R]. Technical Report C-1997-23, Department of Computer Science, University of Helsinki, 1997.

[3] Luhn H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. IBM Journal of Research and Development, 1957, 1(4): 309-317.

[4] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research Development, 1958, 2(2): 159-165.

[5] Francis W N, Kučera H, Mackie A W. Frequency Analysis of English Usage[J]. Frequency Analysis of English Usage Lexicon & Grammar, 1982, 18: 64-70.

[6] Frakes W B, Baeza-Yates R. Information Retrieval: Data Structures and Algorithms[M]. Prentice-Hall, Inc. ,1992.

[7] Lo T W, He B, Ounis I. Automatically Building a Stopword List for an Information Retrieval System[J]. Journal of Digital Information Management, 2005, 3(1): 3-8.

[8] 江兆中. 基于语境和停用词驱动的中文自动分词研究[D]. 合肥：合肥工业大学，2010. (Jiang Zhaozhong. Chinese Words Segmentation Based on Context and Stopwords[D]. Hefei: Hefei University of Technology, 2010.)

[9] 熊文新，宋柔. 信息检索用户查询语句的停用词过滤[J]. 计算机工程，2007，33(6): 195-197. (Xiong Wenxin, Song Rou. Removal of Stop Word in Users' Request for Information Retrieval[J]. Computer Engineering, 2007, 33(6): 195-197.)

[10] 周钦强，孙炳达，王义. 文本自动分类系统文本预处理方法的研究[J]. 计算机应用研究，2005(2): 85-86. (Zhou Qinqiang, Sun Bingda, Wang Yi. Study on New Pretreatment Method for Chinese Text Classification System[J]. Application Research of Computers, 2005(2): 85-86.)

[11] Yang B Y, Pedersen J O. A Comparative Study on Feature [C]//Proceedings of International Conference on Machine Learning. 2010.

[12] Silva C, Ribeiro B. The Importance of Stop Word Removal on Recall Values in Text Categorization[C]// Proceedings of the International Joint Conference on Neural Networks.2003, 3: 20-24.

[13] Tomov D T. Some Critical Remarks on the Stop Word Lists of ISI Publications[J]. Journal of Documentation, 2001, 57(6): 798-808.

[14] 化柏林. 知识抽取中的停用词处理技术[J]. 现代图书情报技术，2007(8): 48-51. (Hua Bolin, Stop-Word Processing Technique in Knowledge Extraction[J]. New Technology of Library and Information Service, 2007(8): 48-51.)

[15] Van Rijsbergen C J. Information Retrieval[M]. London: Butterworths, 1975.

[16] Fox C. A Stop List for General Text[J]. ACM SIGIR Forum, 1990, 24(1-2): 19-21.

[17] 陈欣，张菁，李晓光，等. 一种面向中文敏感网页识别的文本分类方法[J]. 测控技术，2011,30(5): 27-31. (Chen Xin, Zhang Jing, Li Xiaoguang, et al. A Text Classification Method

for Chinese Pornographic Web Recognition[J]. Measurement & Control Technology, 2011,30(5): 27-31.)

[18] 顾益军, 樊孝忠, 王建华, 等. 中文停用词表的自动选取[J]. 北京理工大学学报, 2005, 25(4): 337-340. (Gu Yijun, Fan Xiaozhong, Wang Jianhua, et al. Automatic Selection of Chinese Stoplist[J]. Transactions of Beijing Institute of Technology, 2005, 25(4): 337-340.)

[19] 崔彩霞. 停用词的选取对文本分类效果的影响研究[J]. 太原师范学院学报:自然科学版, 2008, 7(4): 91-93. (Cui Caixia. Research on the Effect of Stop Words Selection on Text Categorization [J]. Journal of Taiyuan Normal University: Natural Science Edition, 2008, 7(4): 91-93.)

[20] Zou F, Wang F L, Deng X, et al. Automatic Construction of Chinese Stop Word List[C] // Proceedings of the International Conference on Applied Computer Science. 2006: 16-18.

[21] 王素格, 魏英杰. 停用词表对中文文本情感分类的影响[J]. 情报学报, 2008, 27(2): 175-179. (Wang Suge, Wei Yingjie. The Influence of Stoplist on the Chinese Text Sentiment Categorization[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 175-179.)

[22] 周姚. 基于云计算的文本挖掘技术研究[D]. 长沙: 国防科学技术大学, 2011. (Zhou Yao. Cloud Computing-based Research on Text Mining Techniques[D]. Changsha: National University of Defense Technology, 2011.)

[23] Makrehchi M, Kamel M S. Automatic Extraction of Domain-Specific Stopwords from Labeled Documents [C] // Proceedings of European Conference on IR Research(ECIR 2008), Glasgow, UK. 2008: 222-233.

[24] 华林森. 中文文本情感分类研究[D]. 重庆: 重庆大学, 2014. (Hua Linsen. Study on Chinese Text Sentiment Classification[D]. Chongqing: Chongqing University, 2014.)

[25] 搜狗实验室. 搜狐新闻数据[DB/OL]. [2016-07-05]. http://www.sogou.com/labs/resource/cs.php. (Sogou Labs. Sohu News Data [DB/OL]. [2016-07-05]. http://www.sogou.com/labs/resource/cs.php.)

[26] 李梅. 改进的 K 均值算法在中文文本聚类中的研究[D]. 合肥: 安徽大学, 2010. (Li Mei. Study of Chinese Text Clustering on Improved K-means Algorithm[D]. Hefei: Anhui University, 2010.)

[27] 黄磊, 伍雁鹏, 朱群峰. 关键词自动提取方法的研究与改进[J]. 计算机科学, 2014, 41(6): 204-207. (Huang Lei, Wu Yanpeng, Zhu Qunfeng. Research and Improvement of TFIDF Text Feature Weighting Method[J]. Computer Science, 2014, 41(6): 204-207.)

[28] 数据堂. 文本分类语料库(复旦)测试语料[DB/OL]. [2016-07-05]. http://www.datatang.com/datares/go.aspx?dataid=615059. (Data Hall. Text Classification Corpus (Fudan) Test Corpus [DB/OL]. [2016-07-05]. http://www.datatang.com/datares/go.aspx?dataid=615059.)

[29] 胡晓辉. 基于团结构的文本分类技术研究[D]. 南昌: 江西师范大学, 2008. (Hu Xiaohui. The Research on Text Classification Based on Clique Model[D]. Nanchang: Jiangxi Normal University, 2008.)

[30] 孙国菊, 张杰. 中文文本分类的特征选取评价[J]. 哈尔滨理工大学学报, 2005, 10(1): 76-78. (Sun Guoju, Zhang Jie.An Evaluation of Feature Selection Methods for Text Categorization[J]. Journal of Harbin University of Science and Technology, 2005, 10(1): 76-78.)

[31] 数据堂. 中文文本分类语料[DB/OL]. [2016-07-05]. http://www.datatang.com/data/11971/. (Data Hall. Chinese Text Categorization Corpus [DB/OL]. [2016-07-05]. http://www.datatang.com/data/11971/.)

[32] 数据堂. 停用词集合[DB/OL]. [2016-07-05]. http://www.datatang.com/data/19300/. (Data Hall. Stop Words Set [DB/OL]. [2016-07-05]. http://www.datatang.com/data/19300/.)

[33] 于娟, 尹积栋, 费庶. 基于句法结构分析的同义词识别方法研究[J]. 现代图书情报技术, 2013(9): 35-40. (Yu Juan, Yin Jidong, Fei Shu. Identifying Synonyms Based on Sentence Structure Analysis[J]. New Technology of Library and Information Service, 2013(9): 35-40.)

[34] 费洪晓, 康松林, 朱小娟, 等. 基于词频统计的中文分词的研究[J]. 计算机工程与应用, 2005, 41(7): 67-68. (Fei Hongxiao, Kang Songlin, Zhu Xiaojuan, et al. Chinese Word Segmentation Research Based on Statistic the Frequency of the Word[J]. Computer Engineering and Applications, 2005, 41(7): 67-68.)

作者贡献声明:

官琴: 数据筛选, 进行实验, 论文起草;

邓三鸿, 王昊: 提出研究思路, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: mf1614036@smail.nju.edu.cn。

[1] 官琴, 邓三鸿, 王昊. SohuData.zip. 搜狐新闻数据.

[2] 官琴, 邓三鸿, 王昊. FudanData.zip. 文本分类语料库(复旦)测试语料.

[3] 官琴, 邓三鸿, 王昊. ChineseTextData.zip. 中文文本分类语料.
[4] 官琴, 邓三鸿, 王昊. Stopwords.zip. 停用词集合.

收稿日期: 2016-12-05
收修改稿日期: 2016-12-25

Chinese Stopwords for Text Clustering: A Comparative Study

Guan Qin　Deng Sanhong　Wang Hao
(School of Information Management, Nanjing University, Nanjing 210023, China)
(Jiangsu Key Lab of Data Engineering and Knowledge Service, Nanjing 210023, China)

Abstract: [**Objective**] This paper compares and analyzes the impacts of stopwords on textual data processing, aiming to improve the construction and use of stopwords. [**Methods**] We obtained stopword lists from Baidu Search Engine, Harbin Institute of Technology and the Machine Learning Laboratory of Sichuan University for this study. First, we processed text message with the stopword lists and Chinese word segmentation technique, the TF-IDF feature evaluation function and the VSM vector model. Secondly, we analysed the texts with the K-means algorithm to calculate the P, R and F1 values. [**Results**] Different stopword lists posed various effects to the text data processing tasks. The length of the list and the content structure of the texts directly influenced the clustering results. More importantly, the two-character stopwords was the biggest factor. [**Limitations**] The text types and quantity were limited. More research is needed to analyze the text with different types of stop words. [**Conclusions**] Stopword list poses significant impacts on text clustering, thus, it is extremely important to build or choose the appropriate Chinese stopword list. However, excessively increasing the number of stop words might not always improve the clustering results.

Keywords: Text Clustering　Stopword List　K-means